

ORIGINAL ARTICLE

Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study

Marie-Hélène Metzger^{1,2} | Nastassia Tvardik¹ | Quentin Gicquel^{3,4} | Côme Bouvry^{3,4} | Emmanuel Poulet^{3,5,6} | Véronique Potinet-Pagliaroli⁷

¹ Université Paris 13 Sorbonne Paris Cité, Laboratoire Educations et Pratiques de Santé EA 3412, Bobigny, France

² Hospices Civils de Lyon, Hôpital de la Croix-Rousse, Unité d'hygiène et d'épidémiologie, Lyon, France

³ Université de Lyon, Lyon, France

⁴ CNRS UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France

⁵ Hospices Civils de Lyon, Groupement Hospitalier Edouard Herriot, Psychiatrie des Urgences, Lyon, France

⁶ Université Lyon 1, EA 4615 «SIPAD» (Stimulation in Psychiatric and Addictive Disorders), Lyon, France

⁷ Hospices Civils de Lyon, Hôpital de la Croix-Rousse, Service des urgences, Lyon, France

Correspondence

Marie-Hélène Metzger, Université Paris 13 Sorbonne Paris Cité, Laboratoire Educations et Pratiques de Santé EA 3412, UFR SMBH – Université Paris 13, 74, rue Marcel Cachin, F-93017 Bobigny, France.
Email: metzger.marie-helene@orange.fr

Abstract

The aim of this study was to determine whether an expert system based on automated processing of electronic health records (EHRs) could provide a more accurate estimate of the annual rate of emergency department (ED) visits for suicide attempts in France, as compared to the current national surveillance system based on manual coding by emergency practitioners. A feasibility study was conducted at Lyon University Hospital, using data for all ED patient visits in 2012. After automatic data extraction and pre-processing, including automatic coding of medical free-text through use of the Unified Medical Language System, seven different machine-learning methods were used to classify the reasons for ED visits into “suicide attempts” versus “other reasons”. The performance of these different methods was compared by using the F-measure. In a test sample of 444 patients admitted to the ED in 2012 (98 suicide attempts, 48 cases of suicidal ideation, and 292 controls with no recorded non-fatal suicidal behaviour), the F-measure for automatic detection of suicide attempts ranged from 70.4% to 95.3%. The random forest and naïve Bayes methods performed best. This study demonstrates that machine-learning methods can improve the quality of epidemiological indicators as compared to current national surveillance of suicide attempts.

KEYWORDS

attempted suicide, emergency service, machine learning, natural language processing, population surveillance

1 | INTRODUCTION

In 2000, the World Health Organization (WHO) estimated that one suicide occurred roughly every 40 seconds worldwide (WHO, 2002). Patients who have previously attempted suicide are at a high risk of making a new suicide attempt (12–30%) or completing suicide (1–3%) within the first year. It is therefore crucial to identify patients presenting to emergency departments (EDs) for suicide attempts in order to begin treatment and prevention (Vaiva et al., 2006).

Monitoring of ED visits for suicide attempts is currently organized at the French national level through the OSCOUR network (Chan-Chee & Jezewski-Serra, 2011). This network was created by the French national institute for public health surveillance (InVS) in 2004. In 2012, 378 EDs participated in this network, covering 60% of all ED visits in France. The following data are extracted from the medical

records of each ED patient for the national surveillance network: demographic variables (gender, age) medical data and administrative data (principal diagnosis, secondary diagnoses, severity, mode of transport, etc.). Diagnoses are manually coded by ED physicians using the ICD-10 (WHO, 2010). In the Rhône-Alpes region of France, all ED patient records are computerized. Data extraction and transfer is done automatically via a regional server (OURAL). The data are then transmitted from OURAL to the OSCOUR national server in near real time (daily transmission at 5:00 a.m.). The OSCOUR national health information system was set up primarily to generate health alerts when unusual events (heat waves, flu epidemics, etc.) occur, and also to collect data on specific diseases (infections, asthma, allergies, etc.). The ICD-10 codes used to monitor suicide attempts are X60 to X84.9 (Chan-Chee & Jezewski-Serra, 2011). InVS estimated that 220,000 emergency room visits were related to suicide attempts each year

(Institut National de Veille Sanitaire, 2014). However, no data were available on the quality or thoroughness of the data produced by the network. Data collection was based on ICD-10 diagnostic codes (WHO, 2010) entered manually by emergency physicians at the end of each patient's ED visit, a task that represented an additional workload. The recent development of semantic mining tools could be a more efficient and informative alternative to manual data collection (Hripcsak & Albers, 2013; Pathak, Kho, & Denny, 2013; Patrick, Nguyen, Wang, & Li, 2011; Proux et al., 2009). Indeed, these tools can automatically produce data for epidemiological surveillance but require testing and validation with authentic data (Carrell et al., 2014a, 2014b; Chute, 2014).

The aim of this study was to determine whether automated extraction and processing of computerized ED records containing both structured and unstructured data could yield a more accurate estimate of the annual rate of ED visits for suicide attempts, by comparison with current national surveillance based on manual coding by ED physicians. This pilot study took place at the ED of Lyon University Hospital (Hôpital de la Croix Rousse).

2 | METHODS

2.1 | Definition of non-fatal suicidal behaviour

Suicidal ideation and suicide attempts share several terms in physicians' free-text description of symptoms. Algorithms designed to detect suicide attempts must therefore distinguish between the two situations. The following WHO definitions were used for this study:

A suicide attempt was defined as "non-fatal suicidal behaviour, a self-injury with the desire to end one's life that does not result in death". A broader definition is "an act with a non-fatal outcome, in which an individual deliberately initiates a non-habitual behaviour that, without third-party intervention, would cause self-harm; or deliberately ingests a substance in excess of the prescribed or generally recognized therapeutic dosage, and which aims to realize changes that the subject desired through the expected physical consequences" (Platt et al., 1991).

Suicidal ideation was defined as "thoughts of killing oneself, in varying degrees of intensity and elaboration, or a belief that life is not worth living and a desire not to wake from sleep" (WHO, 2002).

2.2 | Data sources

2.2.1 | Hôpital de la Croix-Rousse, a health care facility of Lyon University Hospital, France

Hôpital de la Croix-Rousse is one of 14 health care facilities that comprise Lyon University Hospital. It has 810 beds distributed among several wards, as well as consultations in general medicine, surgery and obstetrics; rehabilitation and long-term care. From 8:00 a.m. until 7:00 p.m. daily, the ED of this health facility treats adults (over 15 years of age) requiring immediate care.

2.2.2 | Electronic Health Records (EHRs) of the emergency department (ED)

Upon the arrival of a patient in the emergency room, caregivers create an Electronic Health Record (EHR) in the Emergency Medical Record (EMR) module of CRISTAL-NET software, the informatics system used

by Lyon University Hospital. This software is used in 77 EDs in France [corresponding to 11.8% of the total number in France: 655 EDs in 2015 (SAMU-Urgences de France, 2015)]. The health information system permits ongoing data management, the establishment of protocols, coding of both diagnoses (ICD-10) and medical procedures, and production of parameters necessary for the department's activity statistics. The EMR allows emergency physicians to record their findings clearly for colleagues who manage the patient downstream. Data entry takes place in real time. Some input fields are structured [vital signs, ED discharge modality, principal and associated diagnoses (ICD-10)], while others are entered in free text (reason for health care utilization, observations, expert opinion, etc.).

The data selected for this study were those pertaining to patients presenting to the Croix-Rousse ED in 2011 and 2012. Data stored daily in the hospital data warehouse were extracted by using queries written by our Information Systems Department using Business Objects software. Extracted files were in XML format.

2.3 | Construction of training and test sets for automatic detection of suicide attempts

A clinical research associate first screened all computerized records for patients who visited the ED in 2011 ($n = 19,865$ visits) and 2012 ($n = 20,400$ visits), preselecting records that referred to suicidal behaviour ($n = 177$ in 2011, $n = 163$ in 2012). The resulting list was checked by a physician, who excluded 13 cases in 2011 and 15 cases in 2012. The same physician then classified the selected records into three categories of non-fatal suicidal behaviour.

2.3.1 | Suicide attempt (category 1)

- Explicit mention of a suicide attempt in the medical record (chief complaint, clinical observation, specialists' notes) OR
- Relevant context noted in the medical records: any mention of past suicide attempts plus drug overdose.

2.3.2 | Suicidal ideation (category 2)

The earlier-mentioned WHO definition was used (WHO, 2002).

2.3.3 | No mention of suicidal behaviour (category 0)

- No information in the medical records suggesting the patient belonged to one of the other two categories.

On this basis, the number of suicide attempts was 112 in 2011 and 98 in 2012 (see Figure 1). The number of cases of suicidal ideation was 49 in 2011 and 48 in 2012. To limit the workload due to the manual de-identification of the medical records, we conducted a nested case-control study. A sample of patients classified as having no evidence of non-suicidal behaviour (controls) was drawn from the population classified as "non-suicidal behaviour" ($n = 19,704$ in 2011 and $n = 20,254$ in 2012), matched by sex and age at a ratio of 2:1 with patients (cases) classified as having made a suicide attempt (category 1) or having suicidal ideation (category 2 of earlier-mentioned classification). This case-control ratio is frequently used and contained 292 random subjects out of 20,254 controls in 2012. The corresponding sampling fraction was then:

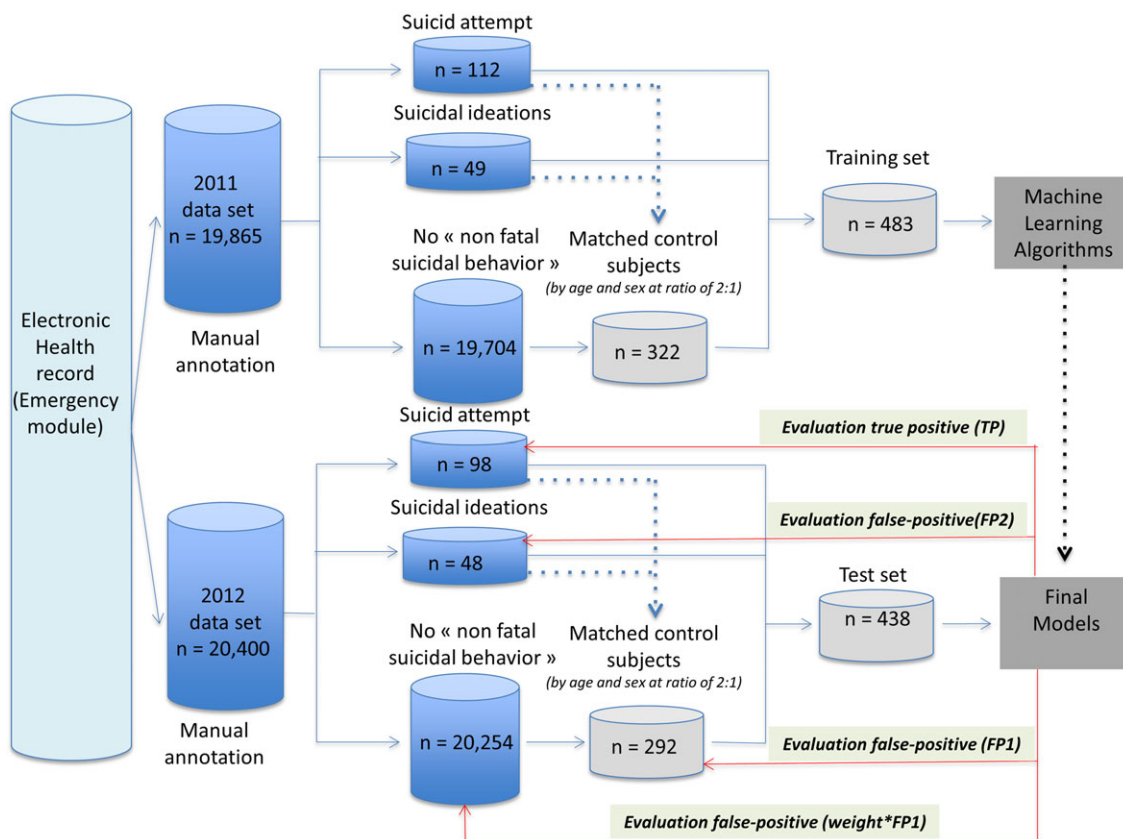


FIGURE 1 Procedure for building the training and test sets

$$\text{Sampling fraction} = \frac{n_2}{n_1} = 0.014417 \quad (1)$$

where n_1 represents the number of patients without a “non-fatal suicide attempt” in the test set ($n = 20,254$); n_2 is the number of matched control subjects in the test set ($n = 292$).

Finally, two data sets were created: the 2011 data (161 cases in categories 1 and 2 and 322 controls in category 0) were used as a training set to establish detection rules for suicide attempts. The 2012 data (146 cases in categories 1 and 2 and 292 controls in category 0) were used as a test set to evaluate the performance of the automated detection system.

2.4 | Pre-processing of the training and test sets

The training data set used 483 medical records totalling 139,663 words, and the test set used 438 medical records totalling 116,372 words. This word count refers to all unstructured fields of the medical records used for this analysis (chief complaint, clinical observation, laboratory tests, diagnostic and therapeutic interventions, typed specialists' notes). The first step of data pre-processing was manual de-identification of all the records. Then relevant sections of the narrative reports (chief complaint, clinical observation, laboratory tests, diagnostic and therapeutic interventions, specialists' notes) were processed with the UrgIndex application, as described in detail elsewhere (Gerber et al., 2011). The application identifies negations in sentences, such as “absence of fever”. The corresponding sentences were then excluded from the process. The application then transforms

medical terms written in natural language into standardized codes. The UrgIndex servlet submits medical terms to a French-language medical multi-terminology indexer (ECMT) developed by CISMef at Rouen University, France (Soualmia, Griffon, Grosjean, & Darmoni, 2011). Five international medical terminologies were selected from the ECMT for this project: ICD-10, the Systematized Nomenclature of Medicine version 3.5 (SNOMED 3.5), the Anatomical, Therapeutic and Chemical (ATC) classification system, Medical Subject Headings (MeSH), and the International Classification of Primary Care (ICPC-2).

To obtain a unique code for each medical term extracted from the medical records, we used the Unified Medical Language System® (UMLS) meta-thesaurus. For each code proposed by the ECMT in one of the five terminologies selected for this study, we sought the CUI (Concept Unique Identifier) for the UMLS meta-thesaurus. The CUI for a meta-thesaurus concept is the identifier to which strings with the same meaning are linked. The CUI allowed us to identify synonymous concepts coded with different codes in the five terminologies used for the same medical concept.

For the training set, the UMLS matching process found 86.5% of the 74,083 ECMT-produced codes among CUI codes. For the test set, the UMLS matching process found 87.3% of the 51 346 ECMT-produced codes among CUI codes.

2.5 | Construction of machine-learning algorithms using the training set

The training set was used to develop machine-learning algorithms, using the features listed in Table 1. Patient classification took place

TABLE 1 Description of the medical record features used for the machine learning algorithms

Feature	Format
Age	Numerical
Gender	Qualitative
Type of admission	Qualitative: structured variable with items to select in the medical record (spontaneous visit, referred by family physician, brought by ambulance, etc.)
Chief complaint	Natural language
Clinical observation	Natural language
Laboratory tests	Natural language
Diagnostic and therapeutic interventions	Natural language
Typed specialists' notes	Natural language
Patient management after the emergency department visit	Qualitative: structured variable (admission to a psychiatric or other ward), discharge with doctor's appointment, self-discharge, return home)

in three steps. The first step consisted of applying seven different machine-learning methods. The methods were applied in parallel. The parameters of the machine learning models were tuned to the training data in order to avoid overfitting due to random fluctuations. The tuning range is described in Table 2.

For each machine learning method, the best classifier was selected by using the F-measure (Chinchor, 1992):

$$F_{\beta} = \frac{(\beta^2 + 1) \text{PPV} * \text{Se}}{(\beta^2 * \text{PPV}) + \text{Se}} \quad (0 \leq \beta \leq +\infty) \quad (2)$$

where PPV is the positive predictive value, Se the sensitivity and $\beta = 2$.

PPV (precision) was defined as the proportion of positive results (suicide attempts) detected automatically by the machine-learning algorithm. Sensitivity (recall) was defined as the proportion of true positives among suicide attempts. The "gold standard" used for these calculations was the patients' manual classification.

The value of β was set to 2 in order to rank classifiers with high sensitivity above classifiers with a high PPV. Indeed, suicide attempts are rare in the general population but need to be detected because of their potential severity, while high PPV is necessary to avoid the workload generated by validation of false positives. The model parameters that yielded the highest F-measures were retained for the second step.

The second step consisted of building two types of classifier (1). The first predicted "non-fatal suicidal behaviour" and the second "suicide attempts" (see Figure 2). This second step was necessary because the vocabulary used in the free-text medical reports to describe suicide attempts is too similar to that used to describe suicidal ideation, and this generates false positives when trying to detect only suicide attempts.

The third step consisted of definitive patient classification, achieved by combining the results of the two types of classifier, using the algorithm shown in Figure 2. Discordances (1) were classified in the "no suicide attempt" category because the error was lower due to the fact that suicide attempts are a rare event. The performance of this algorithm had been validated in a pre-study in which different algorithms were tested with the predictive association rules method (Tvardik, Gicquel, Durand, Potinet-Pagliaroli, & Metzger, 2014). The latter was the first applied in order to follow the quality of data pre-

processing, particularly in terms of the coding quality of medical concepts. The classification algorithm thus obtained was used as the final model for the corresponding machine-learning method.

2.6 | Benchmarking of the final machine-learning models using the test set

The test set ($n = 438$) was used to benchmark the different machine learning algorithms derived from the training set. The F-measure (Equation 1) was calculated for this evaluation. However, because PPV depends on the prevalence of the disease (suicide attempts), it was necessary to estimate the number of false positives, which would have been obtained in the overall study population ($n = 20,400$). This estimation was made by applying the following weight (corresponding to the sampling fraction used for the randomization of the controls in the nested case-control study) to the number of false positives obtained in the test set:

$$\text{Weight}_{\text{False Positives}} = \frac{1}{\text{Sampling Fraction}} = 69.36 \quad (3)$$

The PPV for suicide attempts estimated for the entire ED population and used to estimate the F-measure shown in Table 3 was calculated as follows:

$$\text{PPV}_{\text{Population}} = \frac{\text{TP}}{\text{TP} + \text{FP2} + \text{weight} * \text{FP1}} \quad (4)$$

TP represents the true-positive suicide attempts; FP1 the false-positive suicide attempts detected in the test set (patients without non-fatal suicidal behaviour); FP2 the false-positive suicide attempts detected in the test set (patients with suicidal ideation).

2.7 | Estimation of the annual rate of visits to the Croix-Rousse Hospital ED for suicide attempts, based on machine learning

The annual rate of ED visits for suicide attempts was estimated using each final machine-learning model. The denominator was the annual number of visits to the Croix-Rousse Hospital ED, and the numerator was the annual number of cases detected by the final model.

TABLE 2 Supervised machine learning techniques and corresponding model parameters

Model	R-package	Parameter	Parameter description	Tuning range (suicidal behavior or non fatal suicidal behavior)
Predictive association rules	R arules package, "apriori function" version 1.0–14 (Agrawal, Imieliński, & Swami, 1993; Hahsler & Grün, 2005; Liu, Hsu, & Ma, 1998)s	Support; Confidence	The support $\text{supp}(X)$ of an itemset X is defined as the proportion of transactions in the data set which contain the itemset. The confidence of a rule is defined: $\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y) / \text{supp}(X)$.	$\text{Supp}(X) \in \{0.025; 0.0375; 0.0625; 0.075\}$ $\text{Conf}(X \Rightarrow Y) \in \{0.90; 0.95\}$
Decision trees	RWeka package, Weka classifier trees, "J48 function", version 0.4–18; C4.5 algorithm (Hornik et al., 2015)	Complexity parameter Minsplit	Complexity parameter (cp) specifies the minimum amount of improvement that must be made in order for a split to take place Minsplit: minimum number of observations that must exist in a node in order for a split to be attempted	$cp \in \{0.15; 0.25\}$ $\text{Minsplit} \in \{2; 5; 10; 15; 20\}$
Neural networks	nnet R package, "nnet function" version 7.3–8 (Ripley, 2015)	Decay Size	Decay: parameter for weight decay Size: number of units in the hidden layer. Can be zero if there are skip-layer units	Decay (default value): 0 Size: 0
Logistic regression	R version "386 3.0.1", package "stats"; "GLM function"	—	—	—
Random forest	randomForest R package, "randomForest function" version 4.6–7 (Breiman, 2001; Liaw & Wiener, 2015)	nodesize ntree mtry	nodesize: Minimum size of terminal nodes. ntree: Number of trees to grow mtry: Number of variables randomly sampled as candidates at each split	nodesize $\in \{5; 10; 20\}$ ntree $\in \{25; 51; 75; 101\}$ mtry $\in \{10; 30; 50; 100; 200\}$
Support Vector Machine	e1071 R package, "SVM functions" 0.4–18 (Meyer et al., 2015)	Cost parameter Kernel smoothing parameter	Kernel: the kernel used in training and predicting Cost: cost of constraints violation (default: 1)—it is the 'C'-constant of the regularization term in the Lagrange formulation. Degree: parameter needed for kernel of type polynomial (default: 3) Gamma: parameter needed for all kernels except linear (default: $1/(\text{data dimension})$) Coef0: parameter needed for kernels of type polynomial and sigmoid (default: 0)	Kernel: polynomial; linear; radial; sigmoid Coef0 $\in \{0; 1\}$ Degree $\in \{2; 3; 5; 10\}$ Cost $\in \{0.1; 0.05\}$ Gamma $\in \{0.0008; 0.01; 0.1; 0.5; 1\}$
Naïve Bayes	bnlearn R package, "naive.bayes function" version 3.4 (Scutari, 2010)	—	—	—

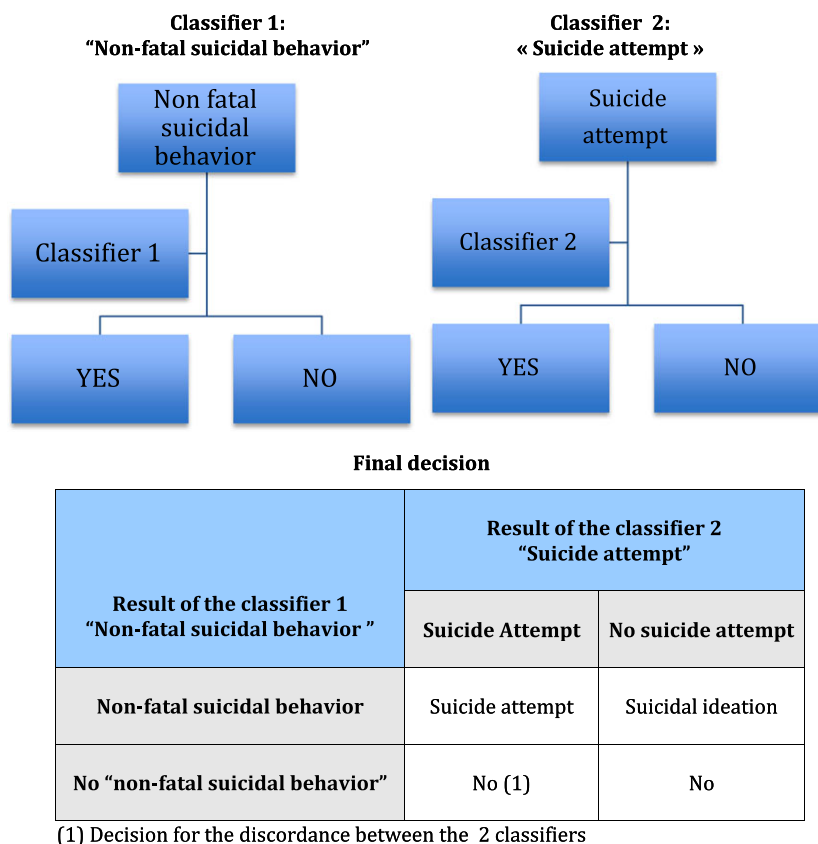


FIGURE 2 Detection algorithm for suicide attempts

TABLE 3 Performance of the different supervised machine learning techniques for the detection of suicide attempts

Supervised machine learning technique	True suicide attempts detected	False-positive suicide attempts detected		Sensitivity (%)	Positive predictive value (%)	F-Measure
	(n)	Suicidal ideations (n) ^a	Other false positives (n) ^b			
Random Forest	94	7	0	95.9	93.1	95.3
Naive Bayes	93	3	0	94.9	96.9	95.3
Support Vector Machines	88	7	0	89.8	92.6	90.4
Predictive Association rules	88	8	0	89.8	91.7	90.2
Decision Trees	85	8	0	86.7	91.4	87.6
Neural Networks	77	9	139	78.6	89.5	80.2
Logistic regression	88	6	139	89.8	93.6	70.4

^aPatient detected as "suicide attempt" by the supervised machine learning technique, but in reality a patient with "suicidal ideation".

^bThe sample dataset comprised 98 cases (patients with true suicide attempts) and 292 controls (patients without non-fatal suicidal behaviour); the number of false-positive suicide attempts was estimated among all emergency department visits without non-fatal suicidal behaviour (n = 20,254)

2.8 | Comparison of the annual rate of visits to the Croix-Rousse Hospital ED for suicide attempts obtained with the machine learning techniques, and that estimated in the national surveillance network

We compared the annual rate of ED visits for suicide attempts based on data transmitted by the Croix-Rousse ED to the OURAL regional server with that obtained using our machine learning techniques.

2.9 | Ethical approval

In keeping with French law, the study was recorded in the processing and applications register of Lyon University Hospital (n° 13–160). Formal patient consent was not required for this type of study.

3 | RESULTS

3.1 | Characteristics of the study population

The number of patients over 15 years of age who presented to the ED was 15,817 (19,865 visits) in 2011 and 16,271 (20,400 visits) in 2012. Based on manual classification of the computer records, the number of suicide attempts was 112 (42 men and 70 women) in 2011 and 98 (37 men and 61 women) in 2012. The number of cases of suicidal ideation was 49 in 2011 (15 men and 34 women) and 48 in 2012 (21 men and 27 women). The age distribution of patients presenting to the ED, both overall and for suicide attempts and suicidal ideation, is shown in Figure 3, separately for men and women.

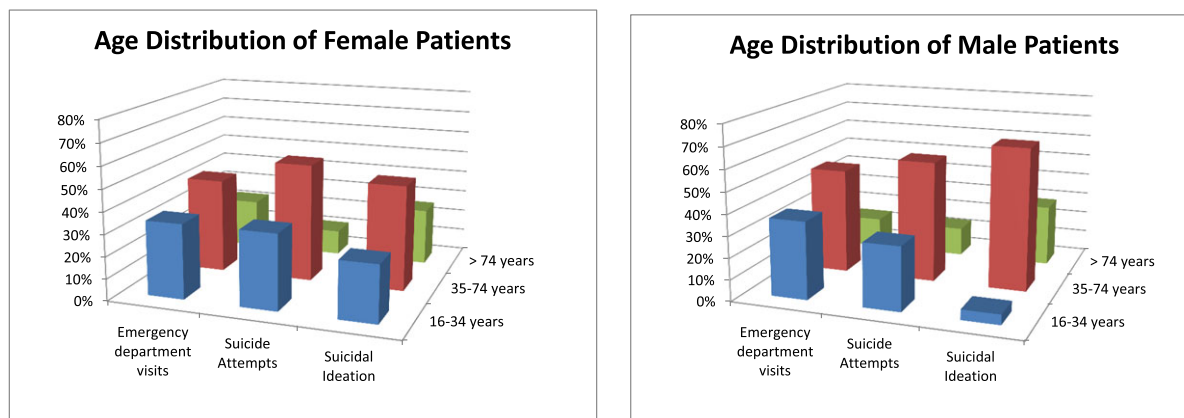


FIGURE 3 Age distribution of male and female patients presenting to the emergency department for suicide attempts and suicidal ideation (Hôpital de la Croix Rousse, Lyon, France, 2012)

3.2 | Benchmarking of the final machine learning models

The performance of the different methods varied widely, with F-measures ranging from 70.4 to 95.3 (Table 3). The methods with the best F-measures were the random forest method and the naïve Bayes classifier, with very close performance. The difference between these two methods concerned the detection of suicidal ideation. The number of cases of suicidal ideation (false-positive suicide attempts) detected by the random forest method was higher than the number detected by the naïve Bayes classifier (respectively seven and three cases).

3.3 | Estimation of the annual number of ED visits for suicide attempts, using machine-learning techniques

Table 4 shows the estimated annual rates of ED visits for suicide attempts obtained with the final machine learning models. Of the seven machine-learning methods evaluated in this work, five yielded estimates close to that of the gold standard method and would thus be valuable for epidemiological surveillance of suicide attempts. Neural networks and logistic regression methods overestimated the annual rate because they included many false positives (other than true suicidal ideation).

3.4 | Comparison of the annual rate of visits to the Croix-Rousse Hospital ED for suicide attempts estimated with the machine learning methods versus the national surveillance network

Of the 98 Croix-Rousse Hospital ED visits related to suicide attempts in 2012, only 15 cases were recorded in the OSCOUR network, leading to significant underestimation of the rate of ED visits for suicide attempts (0.74 ‰ with the OSCOUR surveillance network, versus 4.80 ‰ with the “manual classification” method).

4 | DISCUSSION

The different machine learning methods tested here showed marked differences in performance. The best two methods in terms of the F-measure were the random forest and naïve Bayes approaches. Our results are consistent with those of other studies comparing different methods of data analysis for disease prediction. For example, Ogunyemi, Teklehaimanot, Patty, Moran, and George (2013) showed that, for predicting diabetic retinopathy from a dataset generated in US urban safety-net settings, Bayesian classifiers (sensitivity: 28.5%; PPV: 57.8%) performed better than neural networks (sensitivity: 17%; PPV: 36.7%). Machine-learning approaches are increasingly used

TABLE 4 Estimated annual rate of emergency department visits for suicide attempts in 2012 according to the detection algorithm (total number of emergency department visits in the year 2012: 20,400)

	Estimated number of suicide attempts (n)	Estimated annual rate of emergency department visits for suicide attempts (‰)
Manual Classification (gold standard)	98	4.8
Random Forest	101	5.0
Naive Bayes	96	4.7
Support Vector Machines	95	4.7
Predictive Association Rules	96	4.7
Decision Trees	93	4.6
Neural Networks	225	11.0
Logistic Regression	233	11.4

for cancer prediction and prognostication (Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015)

Recent publications describe the use of such methods for the recruitment of patients for genetic and clinical research (Carrell et al., 2014b; Castro et al., 2015). Carrell et al. (2014b) used natural language processing to improve the efficiency of manual chart abstraction for the identification of breast cancer recurrence. The detection method was based solely on heuristic rule algorithms (Gicquel et al., 2015). Castro et al. (2015) showed that semi-automated mining of EHRs could be used to identify bipolar-disorder patients and healthy control subjects with high specificity (0.53) and high PPV (0.85), as compared with diagnostic interviews. Performance was lower in this latter study than in ours, but the method was different, mixing clinician-defined heuristic rules (in particular for identifying terms consistent with the diagnosis of bipolar disorder) and supervised methods. Our approach was based solely on supervised methods. Furthermore, our work is not intended to reduce manual abstraction efforts in epidemiological studies but rather to use new automatic methods to directly estimate disease incidence. This study shows that the annual rate of ED visits for suicide attempts in Croix-Rousse Hospital is massively underestimated by the manual coding system used by the current national surveillance network. This manual coding also represents an extra workload for ED physicians and has limited epidemiological utility given its poor quality. Our study suggests that it could be replaced by text-mining methods for epidemiological surveillance of suicide attempts.

To limit the workload due to the manual de-identification of the medical records, it was not possible to use the entire dataset (19,865 medical records in 2011 and 20,400 in 2012). We used then a sampling methodology used currently in epidemiology for nested case-control studies. As described by Biesheuvel et al. (2008) diagnostic accuracy estimates derived from a nested case-control study, should be virtually identical to a full cohort analysis. We selected all the cases and selected randomly a sample of patients classified as having no evidence of non-suicidal behaviour (controls), matched by sex and age at a ratio 1:2 [a frequently used case-control ratio (Biesheuvel et al., 2008)]. The controls are then selected from the source population and as described in Rothman, Greenland, and Lash (2008, p. 112) "the case-control study is then a cohort study with data missing at random and by design" and the study is said to be "population-based".

This under-sampling methodology has also the advantage to propose a resampling method for the imbalanced datasets (Chawla, 2010). Due to the fact that suicide attempt is a rare event in the population, there is an imbalance in the dataset used for evaluating performances of the different algorithms. The classification categories are not equally represented in the dataset: the "healthy class" is far better represented than the "diseased" class. Class imbalance hampers the performance of standard classification models (Imran, Afroze, Kumar, & Qyser, 2015). Different variants of over- and under-sampling are possible. Chawla (2010, p. 854) reported: "the random under and over sampling methods have their various shortcomings. The random under-sampling method can potentially remove certain important examples and random oversampling can lead to over-fitting". Japkowicz (2000) compared different techniques and noted that both sampling approaches were effective. Using the sophisticated sampling techniques did not give any clear advantage in the domain considered.

More recently, Imran et al. (2015, p. 1289) found that the effects on the bias and variance components of the imbalanced class distribution depends on the learning algorithm and conclude that "more research is needed to investigate how to best combine under-sampling and over-sampling". To our knowledge, there is no consensus at this stage how to proceed. Using evaluation indicators recommended for imbalanced datasets (sensitivity and PPV) (Chawla, 2010), we estimated the PPV in the entire population, multiplying false positive cases found in the control samples by (1/sampling fraction) (Biesheuvel et al., 2008).

This study has certain limitations. First, as our hospital has no psychiatric experts on site, the emergency transport system preferentially refers suicidal patients to the hospitals with psychiatric departments. We are currently planning a multicentre study that will include hospital facilities offering a range of psychiatric services. This will provide more comprehensive detection of suicide attempts including a broader range of indicative free-text terms representing more underlying causes, and coverage of regional variations in the use of these terms. Another limitation is the relatively small number of suicide attempts, which prevented us from classifying the cases more precisely according to the level of uncertainty (certain, probable, possible suicide attempt). A multicentre study with a larger population sample would overcome this limitation. This classification of uncertainty would be interesting for epidemiological surveillance but not for patient management, as all such cases should receive psychiatric assessment. Another limitation is the diversity of suicide methods. Indeed, the majority of suicide attempts identified in this study involved drug overdose, while other methods were rare or absent. The use of data from other emergency services, and a larger number of cases, would help to refine machine-learning algorithms that include the types of suicide attempt. Lastly, UrgIndex cannot extract data on risk factors such as the medical history, current medications, psychiatric disorders, addiction, social isolation, obesity or hypothyroidism. This part of the feasibility study will be revisited using more sophisticated semantic extraction tools such as that developed for the SYNODOS project (Gicquel et al., 2015).

In future, these machine learning methods could feed a national epidemiological surveillance system, integrating, in a single solution, structured and unstructured data extracted from ED computerized records by an expert system based on text and data mining. The applicability is not limited to hospitals with similar electronic records because the system is based on the processing of extracted files in XML format by the hospital information system. Consequently, each hospital information system which can extract the medical record in XML format can participate to this type of analyses. Moreover, they could serve as a real-time decision aid for ED physicians, and this could in turn improve the machine learning algorithms. The expert system should also handle the updating of decision models, taking into account medical case validation. Methods such as incremental learning (Geng & Smith-Miles, 2009) and data stream learning (Gama, 2010) could also be added to the expert system. This type of solution could be integrated with regional patient record-sharing platforms such as SISRA in the Rhône-Alpes region of France (Metzger, Durand, Lallich, Salamon, & Castets, 2012).

Computerized patient records are a very important potential source of information in such diverse areas as medical decision-making, evidence-based medicine, clinical research, epidemiological

surveillance and data/semantic mining (Chapman & Cohen, 2009; McCoy et al., 2015; Seyfried et al., 2009). The tool we propose could be used to monitor temporal and spatial trends in ED visits for suicide attempts, including risk factors, and contribute to a better understanding of the reasons behind suicide attempts. Follow-up of patients who attempt suicide could also identify characteristics predictive of the subsequent suicide risk and thereby help to develop community-based prevention programmes. The data thus collected could also be used to evaluate public health interventions such as the ALGOS case management algorithm, which includes systematic telephone contacts and a "crisis card" (Vaiva et al., 2011), and the SIAM (Suicide Intervention Assisted by Messages) programme, which consists of post-acute preventive text messaging (Berrouguet et al., 2014). The 2002 WHO report on violence and health (WHO, 2002, p. 206) concluded that "there is a great need for rigorous and long-term evaluations of interventions". Development of these tools at the regional level could produce indicators for evaluating these long-term assessments.

5 | CONCLUSION

Suicide prevention is a major public health concern in France but, despite successive national plans implemented since the late 1990s, surveillance remains inadequate. Our prototype data extraction and detection system based on text-mining analysis of ED medical records represents a new source of data on suicide attempts in France, and could replace the existing monitoring system based on manual coding by ED physicians.

ACKNOWLEDGMENTS

This work was supported by Groupement de coopération sanitaire – Système d'information en Santé Rhône Alpes (<https://www.sante-ra.fr>, French regional public health agency).

The authors thank CISMef (<http://www.hetop.eu/hetop/>) for providing free access to the French-language medical multi-terminology indexer, and Ricco Rakotomalala for his thorough methodological reviewing of this manuscript.

DECLARATION OF INTEREST STATEMENT

All authors declare that they have no conflicts of interest.

REFERENCES

- Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *SIGMOD Record*, 22, 207–216.
- Berrouguet, S., Alavi, Z., Vaiva, G., Courtet, P., Baca-García, E., Vidailhet, P., ... Walter, M. (2014). SIAM (Suicide intervention assisted by messages): the development of a post-acute crisis text messaging outreach for suicide prevention. *BMC Psychiatry*, 14(1), 294. DOI:10.1186/s12888-014-0294-8
- Biesheuvel, C. J., Vergouwe, Y., Oudega, R., Hoes, A. W., Grobbee, D. E., & Moons, K. G. (2008). Advantages of the nested case-control design in diagnostic research. *BMC Medical Research Methodology*, 8, 48. DOI:10.1186/1471-2288-8-48
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Carrell, D. S., Halgrim, S., Tran, D. T., Buist, D. S., Chubak, J., Chapman, W. W., ... Savova, G. (2014a). Carrell et al. respond to "Observational research and the EHR". *American Journal of Epidemiology*, 179(6), 762–763. DOI:10.1093/aje/kwt444
- Carrell, D. S., Halgrim, S., Tran, D. T., Buist, D. S., Chubak, J., Chapman, W. W., ... Savova, G. (2014b). Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *American Journal of Epidemiology*, 179(6), 749–758. DOI:10.1093/aje/kwt441
- Castro, V. M., Minnier, J., Murphy, S. N., Kohane, I., Churchill, S. E., Gainer, V., ... Smoller, J. W. (2015). Validation of electronic health record phenotyping of bipolar disorder cases and controls. *American Journal of Psychiatry*, 172(4), 363–372. DOI:10.1176/appi.ajp.2014.14030423
- Chan-Chee, C., & Jezewski-Serra, D. (2011). Hospitalisations pour tentatives de suicide entre 2004 et 2007 en France métropolitaine. *Analyse du PMSI-MCO. Bulletin Épidémiologique Hebdomadaire*, 492–496.
- Chapman, W. W., & Cohen, K. B. (2009). Current issues in biomedical text mining and natural language processing. *Journal of Biomedical Informatics*, 42(5), 757–759. DOI:10.1016/j.jbi.2009.09.001
- Chawla, N. T. (2010). Data Mining for imbalanced datasets: an overview. In O. Maimon, & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (2nd ed.). (pp. 875–886). Berlin: Springer.
- Chinchor, N. (1992) MUC-4 Evaluation Metrics. *Proceedings of the Fourth Message Understanding Conference*, 22–29.
- Chute, C. G. (2014). Invited commentary: observational research in the age of the electronic health record. *American Journal of Epidemiology*, 179(6), 759–761. DOI:10.1093/aje/kwt443
- Gama, J. (2010). *Knowledge Discovery from Data Streams*. London: Chapman and Hall.
- Geng, X., & Smith-Miles, K. (2009). Incremental learning. In *Encyclopedia of Biometrics* (pp. 731–735). Springer: Berlin.
- Gerbier, S., Yarovaya, O., Gicquel, Q., Millet, A. -L., Smaldore, V., Pagliaroli, V., ... Metzger, M. -H. (2011). Evaluation of natural language processing from emergency department computerized medical records for intra-hospital syndromic surveillance. *BMC Medical Informatics and Decision Making*, 11, 50. DOI:10.1186/1472-6947-11-50
- Gicquel, Q., Tvardik, N., Bouvry, C., Kergourlay, I., Bittar, A., Second, F., ... Metzger, M. H. (2015). Annotation methods to develop and evaluate an expert system based on natural language processing in electronic medical records. *Studies in Health Technology and Informatics*, 216, 1067.
- Hahsler, M., & Grün, B. (2005). Arules – a computational environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software*, 14(15), 1–25.
- Hornik, K., Buchta, C., Hothorn, T., Karatzoglou, A., Meyer, D., & Zeileis, A. (2015). *Package "RWeka"* (Vol. 2015). The Comprehensive R Archive Network. Available at website : <https://cran.r-project.org/web/packages/RWeka/RWeka.pdf>.
- Hripcsak, G., & Albers, D. J. (2013). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1), 117–121. DOI:10.1136/amiajnl-2012-001145
- Imran, M., Afroze, M., Kumar, V., & Qyser, A. A. M. (2015). Learning from imbalanced data of diverse strategies with investigation. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(6), 1285–1290.
- Institut National de Veille Sanitaire (2014) Hospitalisations et recours aux urgences pour tentative de suicide en France métropolitaine à partir du PMSI-MCO 2004–2011et d'Oscour® 2007–2011, Saint Maurice, Institut National de Veille Sanitaire.
- Japkowicz, N. (2000) Learning from imbalanced data sets: a comparison of various strategies. *Proceedings of the AAAI/2000 Workshop on Learning from Imbalanced Data Sets*, Austin, TX.
- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8–17.
- Liaw, A., & Wiener, M. (2015) Breiman and Cutler's random forests for classification and regression, volume 2015. The Comprehensive R

- Archive Network. Available at website : <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>.
- Liu, B., Hsu, W., & Ma, Y. (1998) Integrating classification and association rules mining. *International Conference on Knowledge Discovery and Data Mining (KDD'98)*, 80–86.
- McCoy, T. H., Castro, V. M., Rosenfield, H. R., Cagan, A., Kohane, I. S., & Perlis, R. H. (2015). A clinical perspective on the relevance of research domain criteria in electronic health records. *American Journal of Psychiatry*, 172(4), 316–320. DOI:10.1176/appi.ajp.2014.14091177
- Metzger, M. -H., Durand, T., Lallich, S., Salamon, R., & Castets, P. (2012). The use of regional platforms for managing electronic health records for the production of regional public health indicators in France. *BMC Medical Informatics and Decision Making*, 12, 28. DOI:10.1186/1472-6947-12-28
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. -C., & Lin, C.-C. (2015) *Package "e1071"*. The Comprehensive R Archive Network. Available at: <https://cran.r-project.org/web/packages/e1071/e1071.pdf>.
- Ogunyemi, O., Teklehaimanot, S., Patty, L., Moran, E., & George, S. (2013). Evaluating predictive modeling's potential to improve teleretinal screening participation in urban safety net clinics. *Studies in Health Technology and Informatics*, 192, 162–165.
- Pathak, J., Kho, A. N., & Denny, J. C. (2013). Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association*, 20(e2), e206–e211. DOI:10.1136/amiainl-2013-002428
- Patrick, J. D., Nguyen, D. H., Wang, Y., & Li, M. (2011). A knowledge discovery and reuse pipeline for information extraction in clinical notes. *Journal of the American Medical Informatics Association*, 18(5), 574–579. DOI:10.1136/amiainl-2011-000302
- Platt, S., Bille-Brahe, U., Schmidtke, A., Bjerke, T., Crepet, P., De Leo, D., ... Sampaio Faria, J. (1991). Parasuicide in Europe: the WHO/EURO multicentre study on parasuicide. I. Introduction and preliminary analysis for 1989. *Acta Psychiatrica Scandinavica*, 97–104.
- Proux, D., Marchal, P., Segond, F., Kergourlay, I., Darmoni, S., Pereira, S., ... Metzger, M.H. (2009) Natural language processing to detect risk patterns related to hospital acquired infections. *Proceedings of the International Workshop Biomedical Information Extraction*, pp. 35–41, Borovets, Bulgaria.
- Ripley, B. (2015) Feed-Forward Neural Networks with a single hidden layer, and for multinomial log-linear models, volume 2015. CRAN. The Comprehensive R Archive Network. Available at: <https://cran.r-project.org/web/packages/nnet/nnet.pdf>
- Rothman, K. J., Greenland, S., & Lash, T. L. (2008). Case-control studies. In K. J. Rothman, S. Greenland, & T. L. Lash (Eds.), *Modern Epidemiology* (pp. 111–127). PA, Lippincott Williams & Wilkins: Philadelphia.
- SAMU-Urgences de France (2015). *Livre blanc: Organisation de la médecine d'urgence en France: un défi pour l'avenir – Les propositions de SAMU-Urgences de France*. SAMU-Urgences de France: Châteauroux.
- Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3), 1–22. DOI: 10.18637/jss.v035.i03
- Seyfried, L., Hanauer, D. A., Nease, D., Albeiruti, R., Kavanagh, J., & Kales, H. C. (2009). Enhanced identification of eligibility for depression research using an electronic medical record search engine. *International Journal of Medical Informatics*, 78(12), e13–e18.
- Soualmia, L. F., Griffon, N., Grosjean, J., Darmoni, S. J. (2011) Improving Information Retrieval by Meta-Modelling Medical Terminologies. *Proceedings of the 13th conference on Artificial Intelligence in Medicine (AIME)*, Lectures Notes in Artificial Intelligence, 215–219.
- Tvardik, N., Gicquel, Q., Durand, T., Potinet-Pagliaroli, V., & Metzger, M. H. (2014) Use of electronic medical records of the emergency department for an automated epidemiological surveillance of attempted suicide: pilot study in a French University Hospital. *Paper presented at the 20th International Epidemiology Association World Congress of Epidemiology*, Anchorage, AK.
- Vaiva, G., Vaiva, G., Ducrocq, F., Meyer, P., Mathieu, D., Philippe, A., ... Goudemand, M. (2006). Effect of telephone contact on further suicide attempts in patients discharged from an emergency department: randomised controlled study. *BMJ*, 332(7552), 1241–1245. DOI:10.1136/bmj.332.7552.1241
- Vaiva, G., Walter, M., Al Arab, A. S., Courtet, P., Bellivier, F., Demarty, A. L., ... Libersa, C. (2011). ALGOS: the development of a randomized controlled trial testing a case management algorithm designed to reduce suicide risk among suicide attempters. *BMC Psychiatry*, 11, 1. DOI:10.1186/1471-244x-11-1
- World Health Organization (WHO) (2002). *World Report on Violence and Health*. Geneva: WHO.
- World Health Organization (WHO) (2010). *International Statistical Classification of Diseases and Related Health Problems, 10th Revision*. Geneva: WHO.

How to cite this article: Metzger M-H, Tvardik N, Gicquel Q, Bouvry C, Poulet E, Potinet-Pagliaroli V. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. *Int J Methods Psychiatr Res*. 2017;26:e1522. <https://doi.org/10.1002/mpr.1522>